EXHIBIT E

REDACTED

Many-Million Core Processors and their Applications

Joe Bates
Singular Computing LLC
Cambridge, MA

(past)
MIT Media Lab
Carnegie Mellon Computer Science Dept

joe@singularcomputing.com

January 2014

CONSIDER

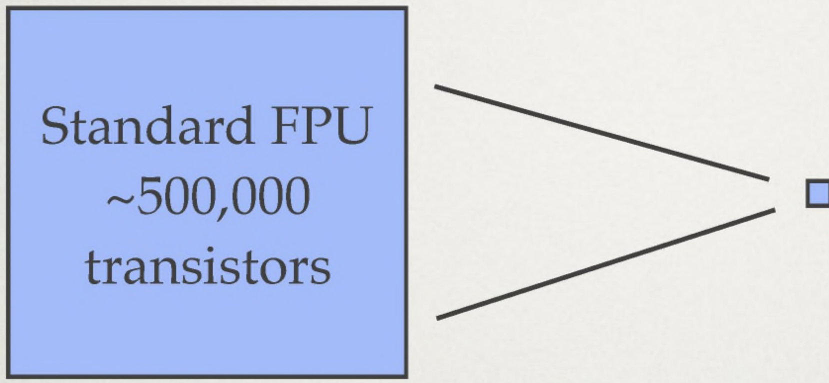
A traditional massively parallel machine,
with floating point arithmetic
that is "99% correct"

(e.g. 1.0 + 1.0 = 1.98 ... 2.02)

"Approximate Computer"

2 *surprises*

- **Surprise #1**: Arithmetic circuit can be <u>unexpectedly</u> small



~100x smaller
standard deterministic
digital cmos

*+ - * / sqrt*
one cycle per op

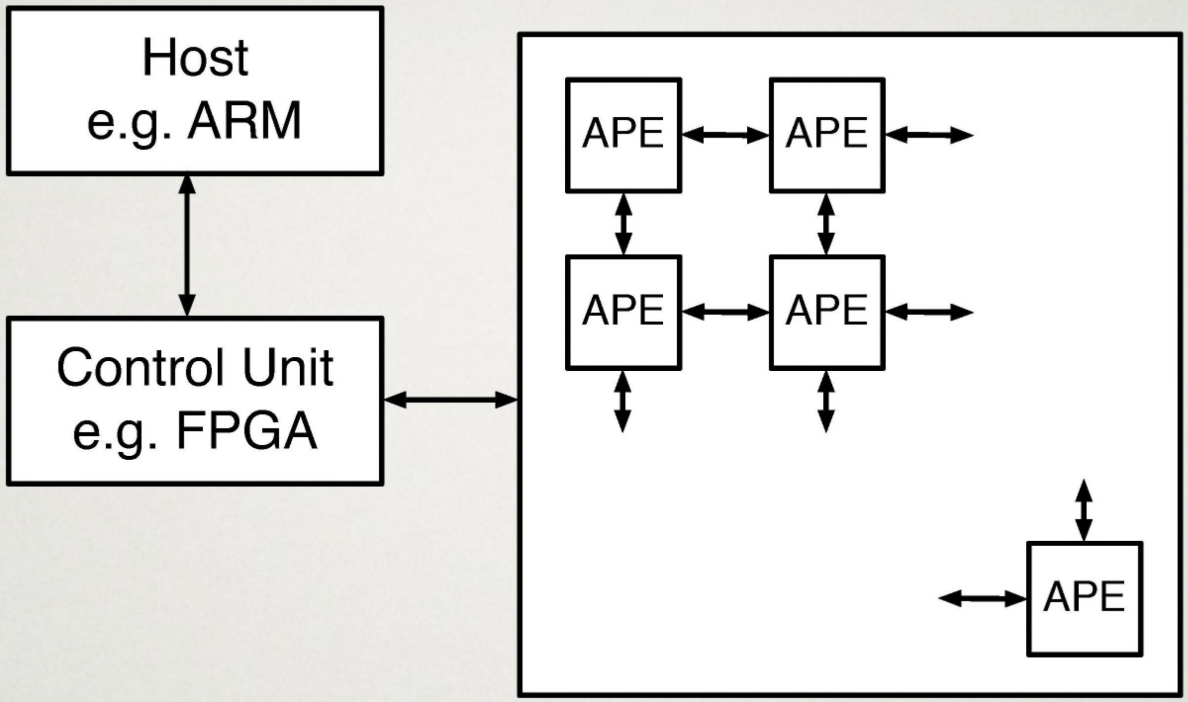
- Add memory, repeat across chip: classic mesh-connected SIMD (e.g. MPP, MasPar, CM, …) with 256,000 cores on chip (1Taflop/watt)

- Usual CPU drives and collaborates with grid
  - fundamentally familiar, single threaded programming model
  - tools, algorithms, papers, proceedings, books from the 80s

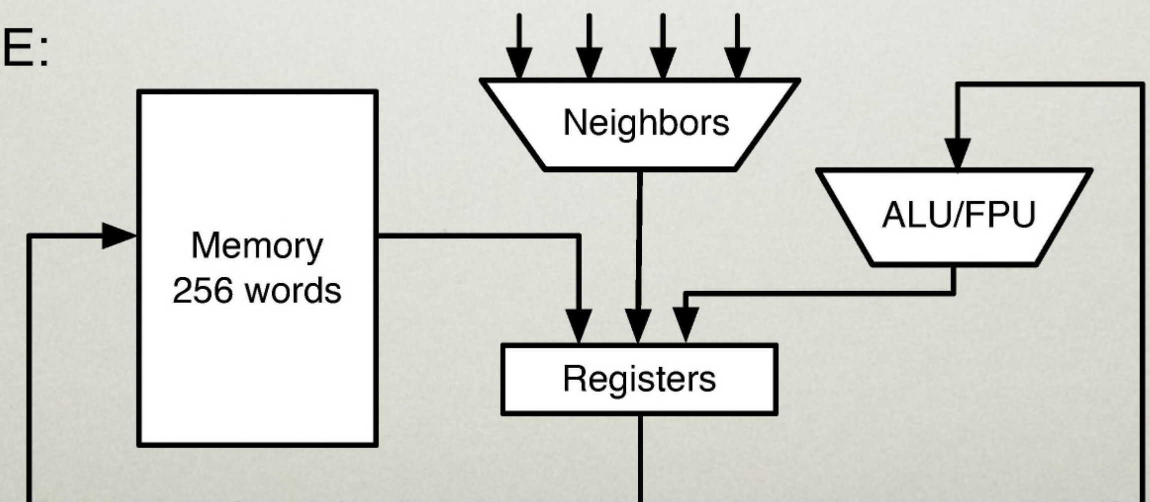
- Mesh naturally scales, e.g.
  embedded: 10K cores, 1 Tflop, sub-watt, few $ in volume
  server: 50M cores, 10 Pflop, single rack

Almost all code here

Kernels here



APE:



approx float exact int + - bool

SIMD: no instruction memory

- Is such a machine useful?

- Massive SIMD well known - hardware constraints
  SIMD, on-chip comm local, small local memory, off-chip bandwidth,
  are manageable for varied important tasks

- But approximate arithmetic is new

- **Surprise #2**: high precision CPU managing low precision workers
  - can yield high precision results (like CPU)
  - but with economics ($, power, size) of the low precision hardware
  - for varied tasks

- *In particular:*
  errors need not compound,
  overall results can be <u>far more accurate</u> than individual operations

SOFTWARE EXAMPLE: K-NEAREST NEIGHBOR

- Workers find <u>several</u> best, CPU picks true best from these:

vector len	% correct if chip finds best 1 / 2 / 3 / 4
200 large, state of art	88.6 / 98.0 / 99.6 / 99.9
800 desired size	79.5 / 93.1 / 97.4 / 99.0

Gaussian (μ=0, σ=1) shown - Uniform(0..1) is similar

Hardware limitations overcome using almost no extra time or energy.

DEBLURRING
(RICHARDSON-LUCY DECONVOLUTION)

original

ieee fp

blurred

approx

Above has 1% error - but collaborating with CPU, can get 100x better (.01%) using 2x time

MIT Media Lab

Motion Detection

Fixed camera - GMM background subtraction

ieee fp

approx

Sometimes no code change needed -- above pixel error rate 0.0014%

ONR / Charles River Analytics

MOTION DETECTION

MOVING CAMERA - ALIGN AND SUBTRACT FRAMES

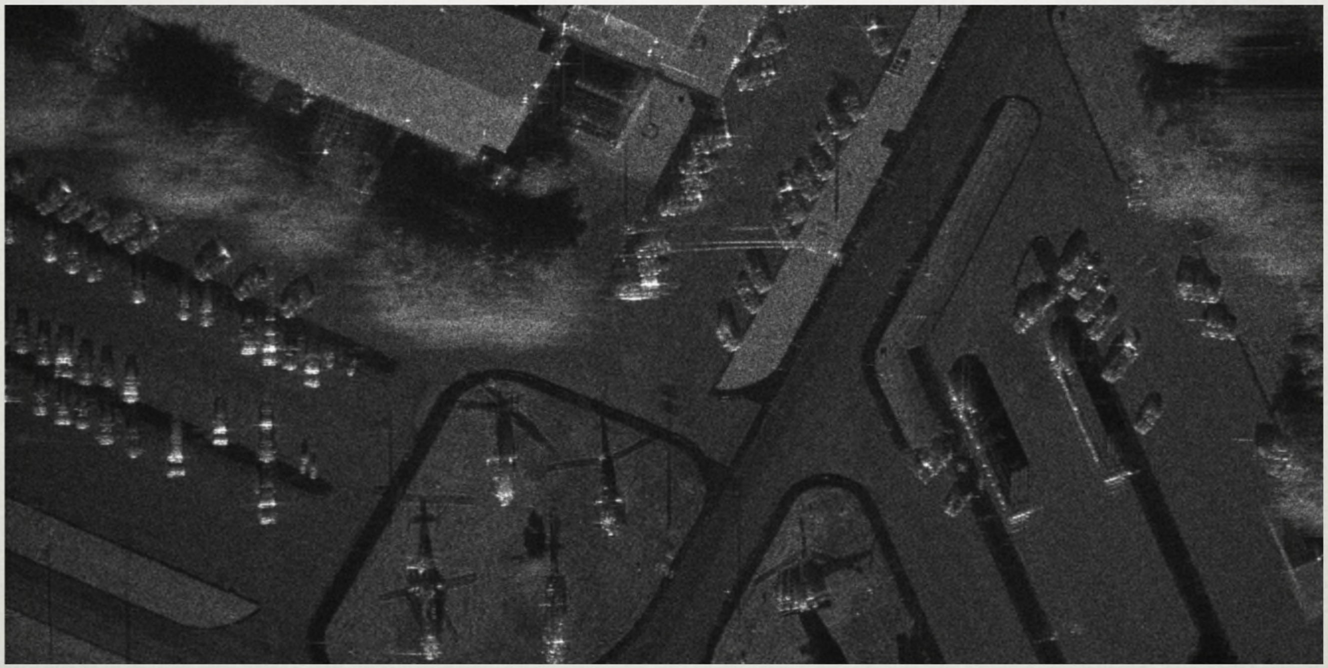
optical flow alignment

brute force patch-match alignment

Simple methods often work well - sophisticated algorithms may come later

Sandia National Labs
SYNTHETIC APERTURE RADAR

- SAR image ⇒ high precision inverse 2D FFT (pseudo phase history)
  ⇒ Singular 2D FFT to form image



- 87% of pixels within +/- 2 (of 255) -- Mean pixel error ~ 0.5%

Don't need to stay away from "math" tasks

ONR / Charles River Analytics
FEATURE BASED TRACKING


CPU only


CPU + Singular

Analysis of system on right
- cpu, bus, accelerator (in emulation) -

CPU quality results,

89x frame rate, 72x less power

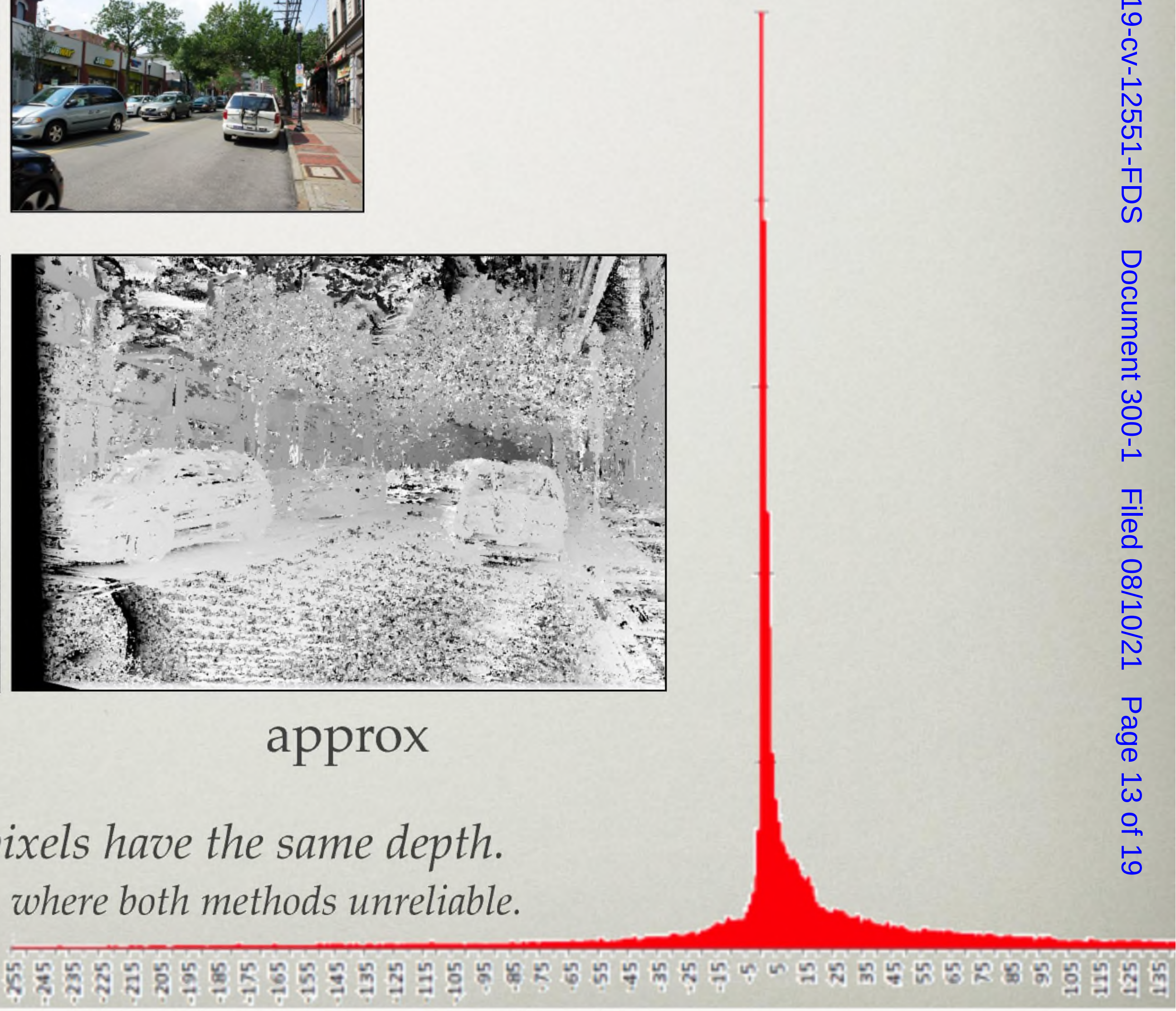
few $ chip

⇒ *6400x better speed/power than CPU alone*

Carnegie Mellon (Takeo Kanade)

DEPTH MAP FROM STEREO VISION

ieee fp approx



Initial results: 78.4% of pixels have the same depth.

Most differences in sky/shadow, where both methods unreliable.

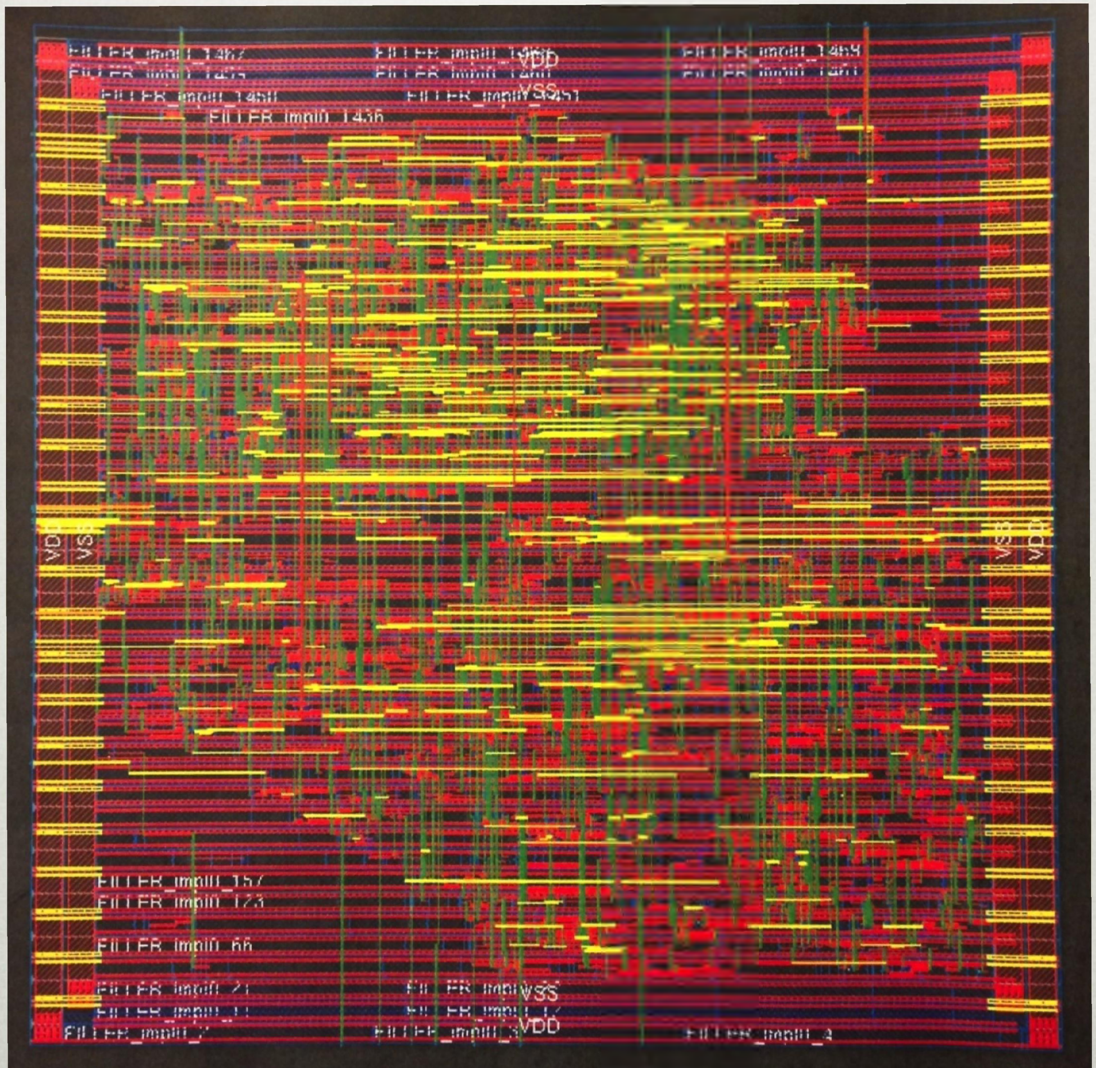
DEEP LEARNING AND OTHER "NEURAL" MODELS

- Multiple efforts in progress:
  internal research, UC Berkeley, MIT, and others
- Believe can do varieties of large scale learning
  making efficient use of the hardware
- e.g. backprop, fully connected, 10 layers, 10K units/layer,
  efficient streaming of large data sets

STATUS

- Program in standard languages, C, Matlab, . . . (build kernels in code, like OpenGL)
- Understandable programming model; varied software surprisingly feasible; approximate hardware need <u>not</u> mean approximate results
- 100x speed per watt (or $) compared to GPU, FPGA, DSP, ARM *(5000x CPU)*
- Building hardware prototypes - DARPA/MTO
  - Cadence, Singular, Intrinsix, MOSIS, GlobalFoundries, BAE, CMU, Berkeley
  - chips: 25mm^2, 40nm, 2500 cores, 150+MHz ⇒ goal ~500Gaflops, 1W
  - systems: ARM + FPGA + 16 Singular ASICs *(40K cores in shoebox)*
  - Use by BAE, CMU, Berkeley, elsewhere ⇒ intending selective open access
- Promising domains:
  image processing/vision
  deep learning (low power embedded forward, or large scale training),
  speech recognition (embedded or at data center, also noise/focus in front end),
  compression, optimization, . . .

BECOMING PHYSICAL . . .



[redacted]

[redacted]

[redacted]